DOCUMENT RESUME

ED 211 595 TM 820 031

AUTHOR	Bejar, Isaac I.; Wingersky, Marilyn S.
TITLE	An Application of Item Response Theory to Equating the Test of Standard Written English.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	College Entrance Examination Board, New York, N.Y.
REPORT NO	CBR-81-8; ETS-RR-81-35
PUB DATE	81
NOTE	34p.
AVAILABLE FROM	College Board Publication Orders, Box 2815, Princeton, NJ 08541 ($4.00).
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	College Entrance Examinations; *Equated Scores; Goodness of Fit; Higher Education; *Latent Trait Theory; Mathematical Models; *Testing Problems
IDENTIFIERS	Item Calibration; *Test of Standard Written English; *Three Parameter Model

ABSTRACT

The study reports a feasibility study for using Item Response Theory (IRT) as a means of equating the Test of Standard Written English (TSWE). The study focused on the possibility of pre-equating, that is, deriving the equating transformation prior to the final administration of the test. The three-parameter logistic model was postulated as the response model and its fit assessed at the item, subscore, and total score level. Minor problems were found at each of these levels but, on the whole, the three-parameter model was found to portray the data well. The adequacy of the equating provided by IRT procedures was investigated in two TSWE forms. It was concluded that pre-equating does not appear to present problems beyond those inherent to IRT-equating. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED211595

College Board Report

No. 81-8

SCOPE OF INTEREST NOTICE
The ERIC Facility has assigned this document for processing to: TM CS
In our judgement, this document is also of interest to the clearinghouses noted to the right. Indexing should reflect their special points of view.

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

An Application of Item Response Theory to Equating the Test of Standard Written English

Isaac I. Bejar
Marilyn S. Wingersky

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
P.K. Hendel
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

An Application of Item Response Theory to Equating the Test of Standard Written English

Isaac I. Bejar

Marilyn S. Wingersky

Educational Testing Service

College Board Report No. 81-8

ETS RR No. 81-35

College Entrance Examination Board, New York, 1981

The authors are grateful to Frederic Lord and Gary Marco who guided various aspects of the project. Also, Linda Cook, Nancy Petersen, and J.B. Sympson offered many substantive and editorial suggestions to improve the report. June Stern was very helpful in the early stages of the project in transmitting some of her knowledge about the most obscure details of the test. Nancy Wright performed many of the "conventional" analyses and provided additional background information.

Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in College Board Reports do not necessarily represent official College Board position or policy.

The College Board is a nonprofit membership organization that provides tests and other educational services for students, schools, and colleges. The membership is composed of more than 2,500 colleges, schools, school systems, and education associations. Representatives of the members serve on the Board of Trustees and advisory councils and committees that consider the programs of the College Board and participate in the determination of its policies and activities.

Additional copies of this report may be obtained from College Board Publication Orders, Box 2815, Princeton, New Jersey 08541. The price is $4.

Copyright © 1981 by College Entrance Examination Board.
All rights reserved.
Printed in the United States of America.

CONTENTS

Abstract iv

Introduction 1
Overview of the Study 1

The Data 1
Description of the TSWE 1
Item Calibration Procedures 3
Step-by-Step Description 4

Fit of the Three-Parameter Model to TSWE Data 6
Evaluation of Fit at the Item Level 6
Evaluation of Fit at the Total Score Level 12
Factorial Structure of TSWE Data 13
Summary 16

Assessment of Pre-equating 16
Equating Based on IRT 17
Results 18

Summary and Conclusions 20

References 23

Appendixes 24

LIST OF TABLES

1 Item and Test Analysis Results for Various TSWE Forms 3

2 Mean and Standard Deviations for Observed and Expected Number-Right Distributions Including and Excluding Students with Omitted Items 13

3 Summary Results of Factor Analysis with Two Item Type Factors 15

4 Summary Conversion Table Comparing Conventional Equating, IRT-equating, and Pre-equating for E7 19

5 Summary Conversion Table Comparing Conventional Equating, IRT-equating, and Pre-equating for E8 19

6 Weighted Mean Squared Difference for Form E7, Using E3, E4, and E5 as the "Old" Form and Three Different Criteria 21

7 Weighted Mean Squared Difference for Form E8, Using E3, E4, and E5 as the "Old" Form and Three Different Criteria 21

LIST OF FIGURES

1 Data Used in the Calibration 5

2 Item-on-Ability Regression for E7 Items with Fixed b Parameter 7

3 Item-on-Ability Regression for E8 Items with Fixed b Parameters 10

4 Observed and Expected Distribution of Number-Right Score Including and Excluding Students with Omitted Responses 14

ABSTRACT

The study reports a feasibility study for using Item Response Theory (IRT) as a means of equating the Test of Standard Written English. The study focused on the possibility of pre-equating, that is, deriving the equating transformation prior to the final administration of the test. The three-parameter logistic model was postulated as the response model and its fit assessed at the item, subscore, and total score level. Minor problems were found at each of these levels but, on the whole, the three-parameter model was found to portray the data well. The adequacy of the equating provided by IRT procedures was investigated in two TSWE forms. It was concluded that pre-equating does not appear to present problems beyond those inherent to IRT-equating.

INTRODUCTION

Equating, in general, refers to the derivation of transformations that map scores on different forms of a test onto a scale in such a way that after transformation the scores on the various forms are comparable. The equating methodology that has been commonly used (see Angoff, 1971) requires that the form being equated first be administered to testees. Since in large-scale testing programs scores are not due back to testees for four to six weeks it would seem that there is ample time to derive the equating transformation. In practice the bulk of the time is consumed by various data processing steps. As a result the equating transformation must be produced in a rather short period of time. Even when no difficulties arise the psychometrician is under considerable pressure.

From this pragmatic point of view, one of the most exciting applications of Item Response Theory (IRT) is pre-equating (see Lord, 1980, Chapter 13). As implied by the name, pre-equating refers to the derivation of the transformation prior to the administration of the form to be equated. This requires that IRT item statistics be available on a common metric for all the items that appear in the final form. The feasibility of implementing pre-equating for the TSWE is the focus of the present study.

Overview of the Study

Whether pre-equating works or not depends on two broad factors. One is the fit of the three-parameter model to TSWE data. Since there is no general procedure for ascertaining fit, several procedures will be used in the hope that collectively they can be more revealing. The second broad factor that may prevent successful pre-equating is lack of "situational" invariance in the item parameter estimates. In practice, pre-equating requires that the final form be assembled from items coming from various pretest forms. This raises the possibility of a context affect on item parameters, which as shown by Yen (1980), can be substantial. The adequacy of pre-equating will be judged on two forms in which these conditions could be simulated using as a criterion scores equated by means of non-IRT procedures.

The next section gives a brief description of the TSWE, as well as the data and calibration procedure used in this study. The following two sections will examine the fit and adequacy of pre-equating, respectively. Recommendations and suggestions for further research will be discussed in the final section.

THE DATA

Description of the TSWE

The TSWE is a 30-minute multiple choice test administered together with the SAT. Its purpose is to help colleges place students in appropriate English Composition courses. It is not recommended as an admission instrument. The test consists of 50 items; items 1-25 and 41-50 are called usage items. The testee is expected to recognize writing that does not follow conventional and standard written English. An example of this type of item is the following:

> Directions: The following sentences contain problems in grammar, usage, diction (choice of words), and idiom.
>
> Some sentences are correct.
> No sentence contains more than one error.

You will find that the error, if there is one, is underlined and lettered. Assume that all other elements of the sentence are correct and cannot be changed. In choosing answers, follow the requirements of standard written English.

If there is an error, select the one underlined part that must be changed in order to make the sentence correct, and blacken the corresponding space on the answer sheet.

If there is no error, mark answer space E.

EXAMPLES:

I. He spoke bluntly (A) and angrily (B) to we (C) spectators. (D) No error (E)

II. He works (A) every day (B) so that he would become financially (C)

independent in his old age. (D) No error (E)

The other 15 items, 26-40, are called sentence correction. In these items, the student is expected to recognize unacceptable usage and to choose the best way of phrasing the sentence. An example of this type of item is the following:

Directions: In each of the following sentences, some part of the sentence or the entire sentence is underlined. Beneath each sentence you will find five ways of phrasing the underlined part. The first of these repeats the original; the other four are different.

If you think the original is better than any of the alternatives, choose answer A; otherwise choose one of the others. Select the best version and blacken the corresponding space on your answer sheet.

This is a test of correctness and effectiveness of expression. In choosing the answer, follow the requirements of standard written English: that is, pay attention to grammar, choice of words, sentence construction, and punctuation. Choose the answer that produces the most effective sentence--clear and exact, without awkwardness or ambiguity. Do not make a choice that changes the meaning of the original sentence.

EXAMPLES:

I. Caroline is studying music because she has always wanted to become it.

(A) it (B) one of them (C) a musician
(D) one in music (E) this

II. Because Mr. Thomas was angry, he spoke in a loud voice.

(A) he spoke (B) and speaking (C) and he speeks
(D) as he spoke (E) he will be speaking

Research on the TSWE has shown it to be a reliable and valid instrument. Table 1 shows some sample statistics for forms E3-E8. As can be seen, standard errors of measurement are about 4.0 and reliabilities are in the upper .80s. For a 30-minute test these figures are satisfactory. Research by Breland (1976) has also provided evidence of the construct validity of scores derived from the TSWE. For example, the correlation between TSWE scores and essay scores is higher than the correlation between SAT verbal scores and essay scores. This is to be expected if indeed the TSWE measures writing ability rather than verbal ability.

TABLE 1. Item and Test Analysis Results for Various TSWE Forms

	Form					
	E3	E4	E5	E6	E7	E8
Admin. date	12/74	2/75	2/75	4/75	11/75	12/75
N	1765	1920	1790	1685	1895	1830
Reliability	.890	.885	.872	.867	.893	.874
SEM (scaled)	3.7	3.8	4.1	4.0	3.6	4.2
Mean R-Bis.	.51	.49	.47	.46	.51	.49
Equated Δ mean	9.2	9.4	9.4	9.6	9.1	8.9

Item Calibration Procedures

Due to the expense involved in item calibration the adequacy of pre-equating was investigated for only two TSWE forms: E7 and E8. As we shall see, however, to obtain item statistics on even two forms is not straightforward. The calibration of a large set of items administered to different samples involves, first, obtaining item parameter estimates on the arbitrary metric defined by each calibration sample, and second, placing all items calibrated on different samples on the same metric.

Parameter estimation. All item parameter estimates used in this report were obtained using the program LOGIST (Wood, Wingersky, and Lord, 1976). This means that the three-parameter-logistic model (Birnbaum, 1968) was the assumed response function. The function of the LOGIST program is to estimate, for each item, the three-item parameters: a (discrimination); b (difficulty); and c (a pseudo-guessing parameter). Unless otherwise indicated, the following constraints were imposed on the estimation: a was restricted between .01 and 1.25; c was held fixed to .15 until stage 2, of step 2. Thereafter a, b, and c are estimated except that c's were held fixed at a constant, $\bar{c}$ estimated by the program for those items with $b-2/a < -2.0$ at the end of stage 3 of step 2 (Wood et al., 1976). The c's for all other items were restricted to a range of .0 to .5.

Putting estimates on the same metric. For no particular reason, other than convenience, the base metric was defined with respect to the December 1974 administration of E3. Two procedures were used to place estimates on the E3 metric. One procedure sets the scale of the items being calibrated by fixing the b estimates of the items previously calibrated. Obviously this requires that the previously calibrated items already be on the desired metric and that they be administered together with the items being calibrated.

The above procedure can be used when previously calibrated items are administered together with uncalibrated items. The second procedure puts item statistics on the desired scale by applying a linear transformation to the item statistics. The procedure requires that the uncalibrated items or new form, be administered by themselves to a random sample of population X, and the previously calibrated items, or old form, also be administered to a random sample of population X. We then calibrate the items separately for the new form and for the old form. The new form is put onto the scale of the old form in the new administration by setting the means and standard deviations

of the abilities equal. We now have two separate estimates of the b parameters for the old form, one from the new administration and one from a previous administration. If the model holds, these estimates are linearly related. A variety of procedures can now be used to derive the linear relationship to transform the b's for the new administration, old form, onto the scale of the previous administration. For example, the mean and standard deviations of the two sets of b estimates can be equated. However, in this report a robust procedure was used. This procedure, adapted by Lord, is explained in Appendix A. Once the transformation is derived it is applied to the a and b parameter estimates of the new form.

Step-by-Step Description

In what follows we will describe each step of the calibration procedure. The following notation will be adopted: TSWE forms are designated by the letter E and a number. To distinguish data from the same form administered to different samples the sample code will precede the TSWE form designation. For example, the parameter estimates designated as W506E3 are obtained on sample W506 responding to the E3 form. The first two characters denote the administration date. It is important to note that samples with the same first two characters are random samples from a given administration. A "P" after a sample designation indicates this set of items consists of pretest items. A "T" at the end of the sample-form designation indicates the parameters have been transformed to the metric defined by W506E3. With this background we now detail the steps of the calibration.

(1) Estimate a, b, c for E3 with the constraints indicated earlier. E3 is the base form and W506 is the base sample.
(2) Estimate a, b, c for E3 on X101 sample.
(3) Derive the transformation X101E3<-->W506E3 using the procedure described in Appendix A.
(4) Estimate a, b, c for E4 based on 1,500 testees from sample X104 and 1,500 from sample X105. These parameters are labeled X104,5E4.
(5) Apply the transformation X101E3<-->W506E3 to the X104,5E4 estimates. The transformed parameters are labeled X104,5E4T.
(6) Estimate a, b, c for pretest items X104P by fixing the b's of the E4 items after transformation, that is, taking the b's from X104,5E4T. This puts the a, b, c estimates for X104P on the W506E3 metric.
(7) Same as Step 6 but for pretest items X105P.
(8) Estimate a, b, c for E5 items on sample X106. The estimates are labeled X106E5.
(9) Apply the X101E3<-->W506E3 transformation to X106E5 estimates. The transformed estimates are labeled X106E5T. (Note that this is legitimate because samples X106 and X101, on which the transformation was derived, are randomly drawn from the same population.)
(10) Estimate a, b, c for pretest items X106P by fixing the b's of E5 to the values in X106E5T.
(11) Estimate a, b, c for E5 on sample X401. Estimates are labeled X401E5.
(12) Derive transformation X401E5<-->X106E5T.
(13) Estimate a, b, c for E7 items on sample X406. Estimates are labeled X406E7.
(14) Apply transformation X401E5<-->X106E5T to X406E7 estimates. Transformed estimates are labeled X406E7T. Again, this is legimate since X401 and X406 are randomly drawn from the same population.
(15) Estimate a, b, c for E5 items on sample X501. Estimates are labeled X501E5.
(16) Derive transformation X501E5<-->X106E5T.
(17) Estimate a, b, c for E8 item on sample X506. Estimates are labeled X506E8.
(18) Apply transformation X501E5<-->X106E5T to X506E8 estimates. Estimates are labeled X506E8T. (See note in step 14.)
(19) Fix the b's of 20 pretested items to the estimates from X104P, X105P, and X106P. Estimate the a, b, and c of the remaining items based on sample Z101; also reestimate the a and c of the 20 pretested items. The parameters are labeled Z101E7.

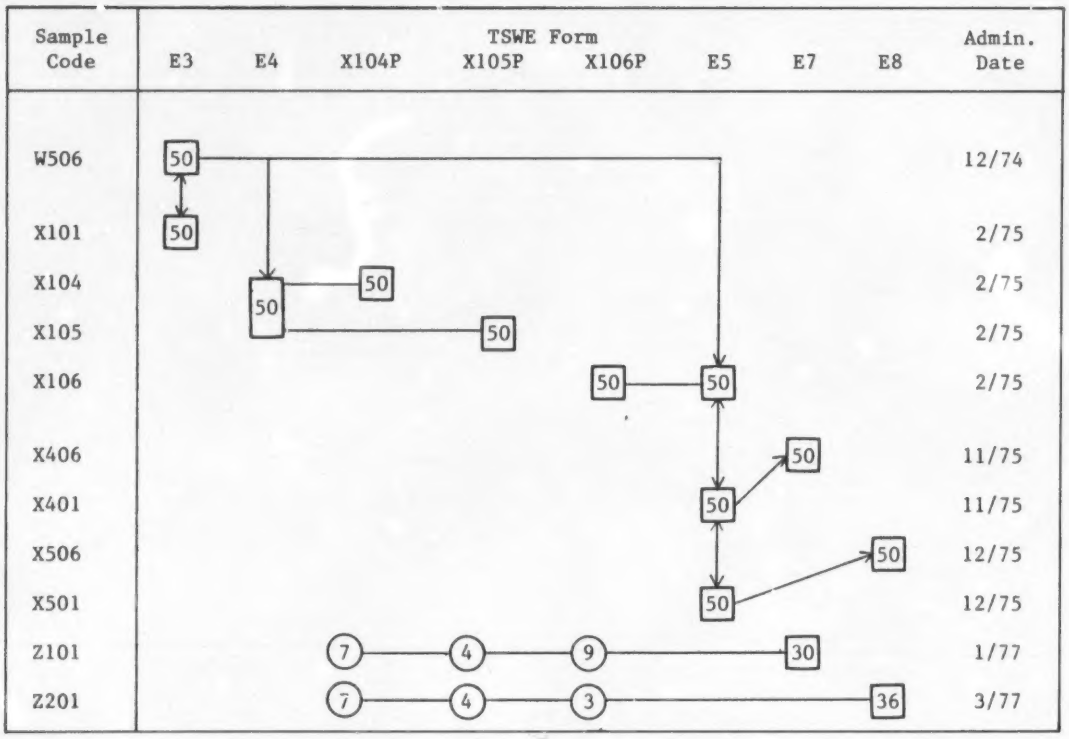


FIGURE 1. Data used in the calibration. The number within a square indicates the number of items for which a, b, and c were estimated. The number within the circle indicates the number of items for which b was fixed and a and c were estimated. A double-headed arrow means the development of a transformation. The single-headed arrow means the application of a transformation. A connecting line without arrows is used to indicate items were administered to the same sample. The forms X104P, X105P, and X106P were pretest forms.

(20) Fix the b's of 14 pretested items to the estimates from X104P, X105P, and X106P. Estimate the a, b, c of the remaining items based on sample Z201. Also re-estimate the a and c of the 14 pretested items. The parameters are labeled Z201E8.

Figure 1 will be helpful in visualizing the calibration procedure. It is also a useful representation of the relationship among the various data sets used in this report. (It should be pointed out that this complex procedure was required to simulate pre-equating conditions.)

Several sets of parameter estimates resulted from the calibration effort. Two additional sets were formed and labeled E7P and E8P. E7P and E8P contain 20 and 14 items, respectively, with the a, b, and c taken from X104P, X105P, and X106P. The remaining 30 and 36 a's, b's, and c's were taken from Z101E7 and Z201E8, respectively.

FIT OF THE THREE-PARAMETER MODEL TO THE TSWE DATA

Conceptually, it seems useful to distinguish between within and between population lack of fit. Within population lack of fit can arise as a result of the violation of the local independence or unidimensionality assumptions. For example responses to certain items in the test may be mediated by a different configuration of cognitive processes. Between population lack of fit on the other hand occurs when different populations respond to the same items with a different configuration of cognitive processes. This may occur, for example, if the demographic composition of the population is different.

These two components of lack of fit will be examined by applying various fit criteria to the data. Specifically, the following procedures were used:

- o Examining the estimated item on ability regression.
- o Contrasting observed and expected distribution of number-right scores.
- o Examining the factorial structure of the TSWE.

Evaluation of Fit at the Item Level

An intuitively appealing way to examine fit is by comparing the observed item on ability regression against the estimated item on ability regression predicted by the model, i.e., the item characteristic curve. The comparison permits a visual assessment of how well the estimated parameters portray the response data for a given item. For the present application plots were constructed as follows, for each item. The estimated item characteristic curve is given by

$$P_i(\theta) = \hat{c}_i + (1 - \hat{c}_i) \left\{ 1 + \exp\left[-1.7\hat{a}_i(\theta - \hat{b}_i)\right] \right\}^{-1}$$

where $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$ are the estimated item parameters for item i, and $P_i(\theta)$ is the probability of answering the item correctly for someone of ability θ.

The observed item on ability regression was computed by dividing the θ into intervals of .4 and grouping students into those intervals based on their $\hat{\theta}$. Within the kth interval the probability of a correct response was computed as

$$P^*_{ik} = \left[NR_k + O_k/A \right] /N$$

where NR_k is the number of testees who answered the item correctly in the kth interval; O_k is the number of students who omitted the item in the kth interval; A is the number of alternatives in the item; and N_k is the total number of testees in the kth interval.

For our purposes it is of most interest to examine the plots for items where the b parameter had been fixed to their pretested value. Figures 2 and 3 shows the plots for the 20 Z101E7 and 14 Z201E8 items for which the b was fixed. The squares constitute the item-on-ability regression; the solid curve is the estimated icc.

The size of the square is proportional to the number of testees in that interval of θ. The asterisk next to the b value indicates the b parameter was fixed to that value. Some items also show an asterisk next to the estimated c. This means that c was fixed to $\bar{c}$, the constant derived earlier (see page 3). For E7, as can be seen, for most of the items the data fit the estimated icc rather well despite the fact the b parameter was estimated from a pretest administration. There are some exceptions, however, including items 6.23, 26, 42, and 46. For E8 most of the items fit the estimated icc with the exception of item 34.

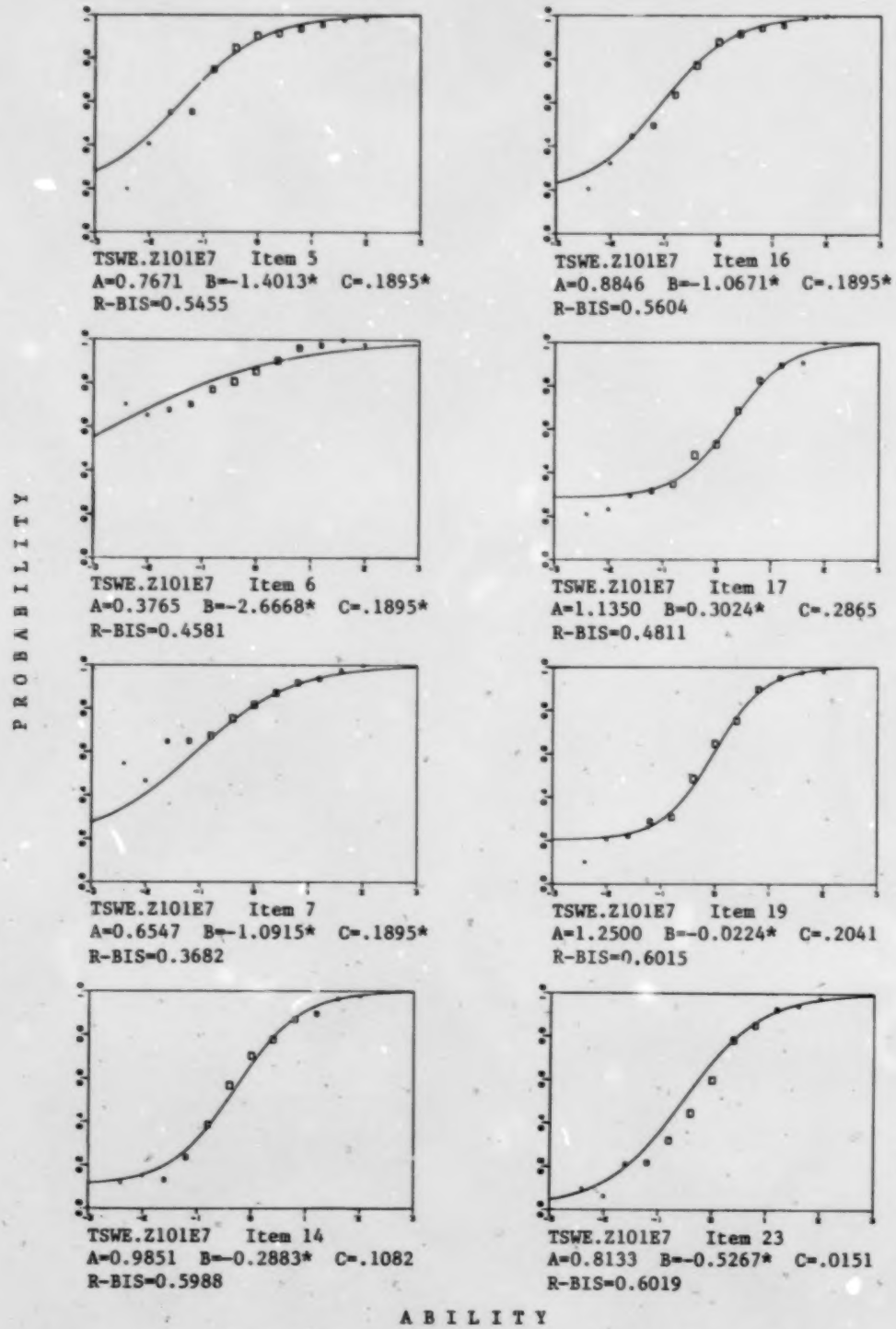


FIGURE 2. Item-on-ability regression for E7 items with fixed b parameter.

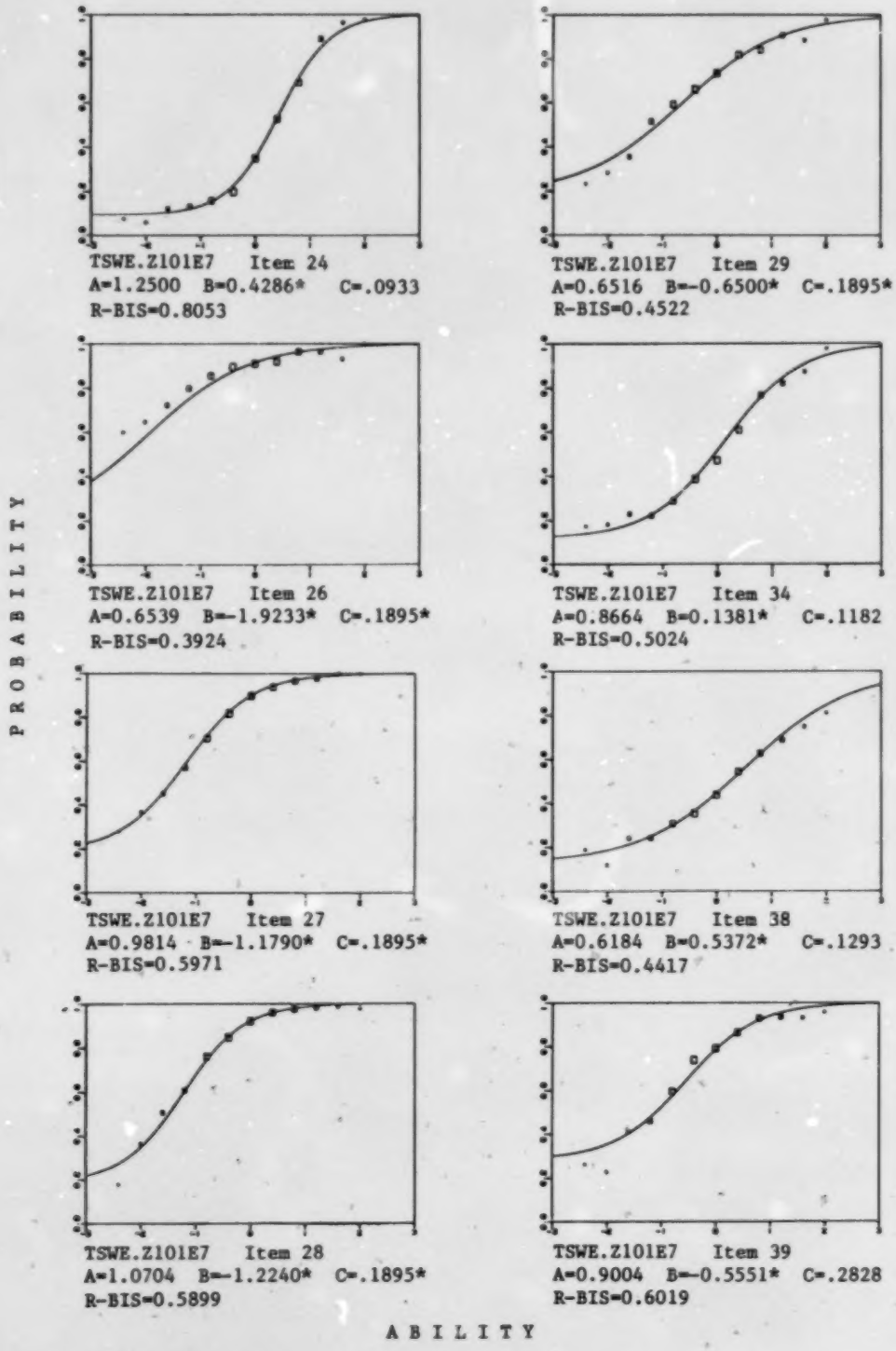


FIGURE 2 (Cont'd)

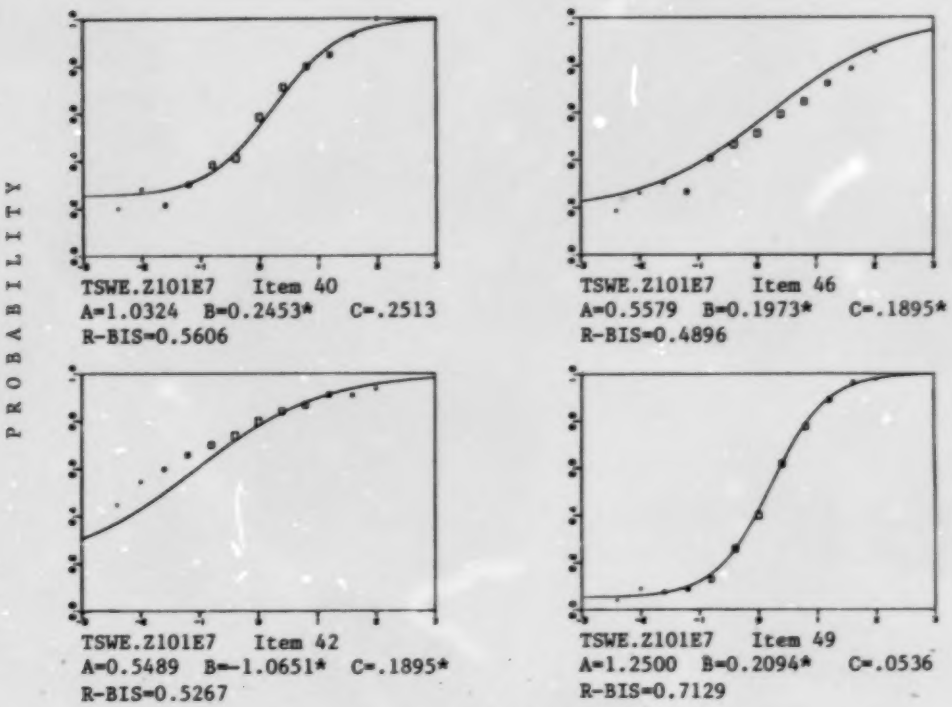


FIGURE 2 (Cont'd)

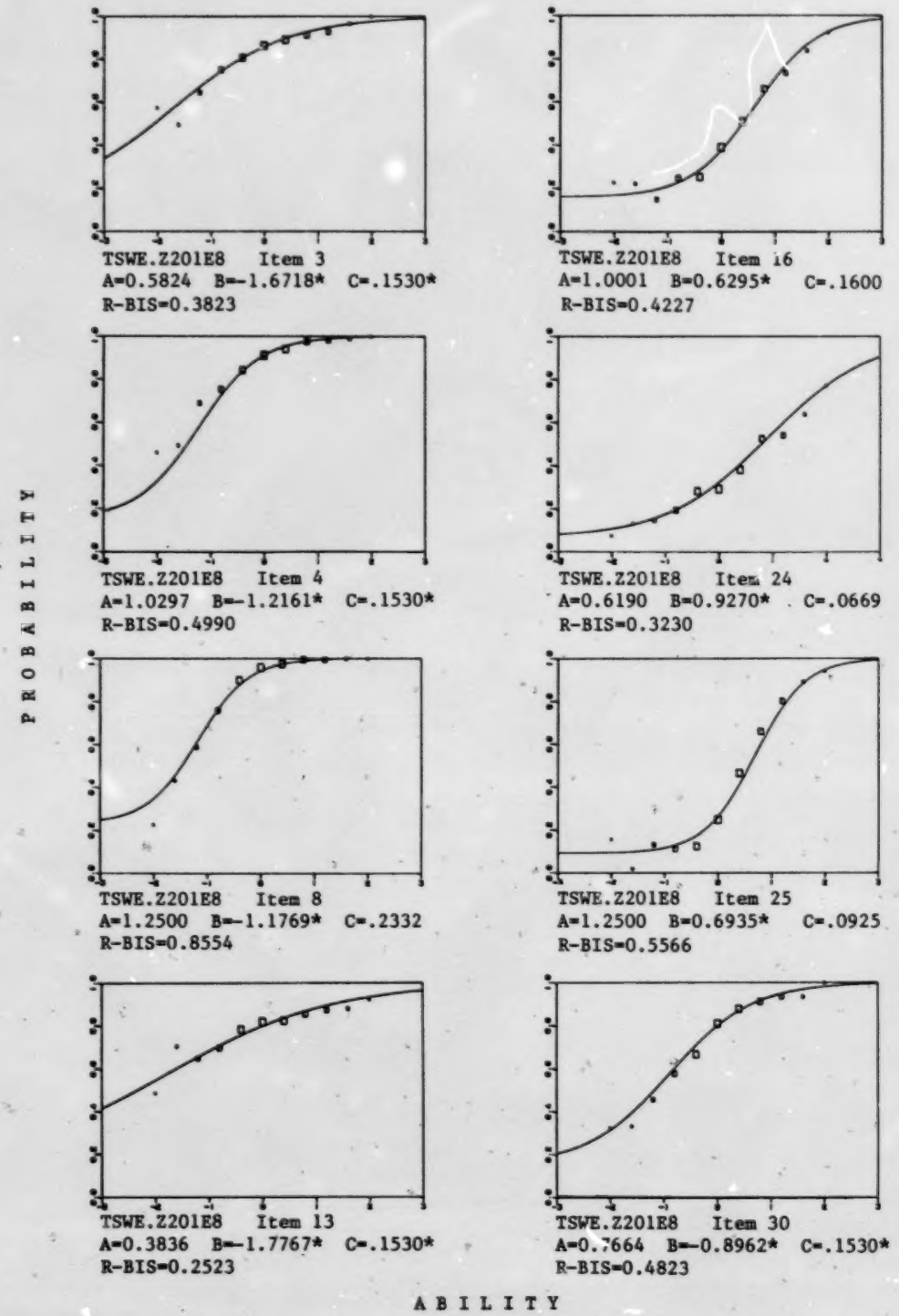


FIGURE 3. Item-on-ability regression for E8 items with fixed b parameters.

10

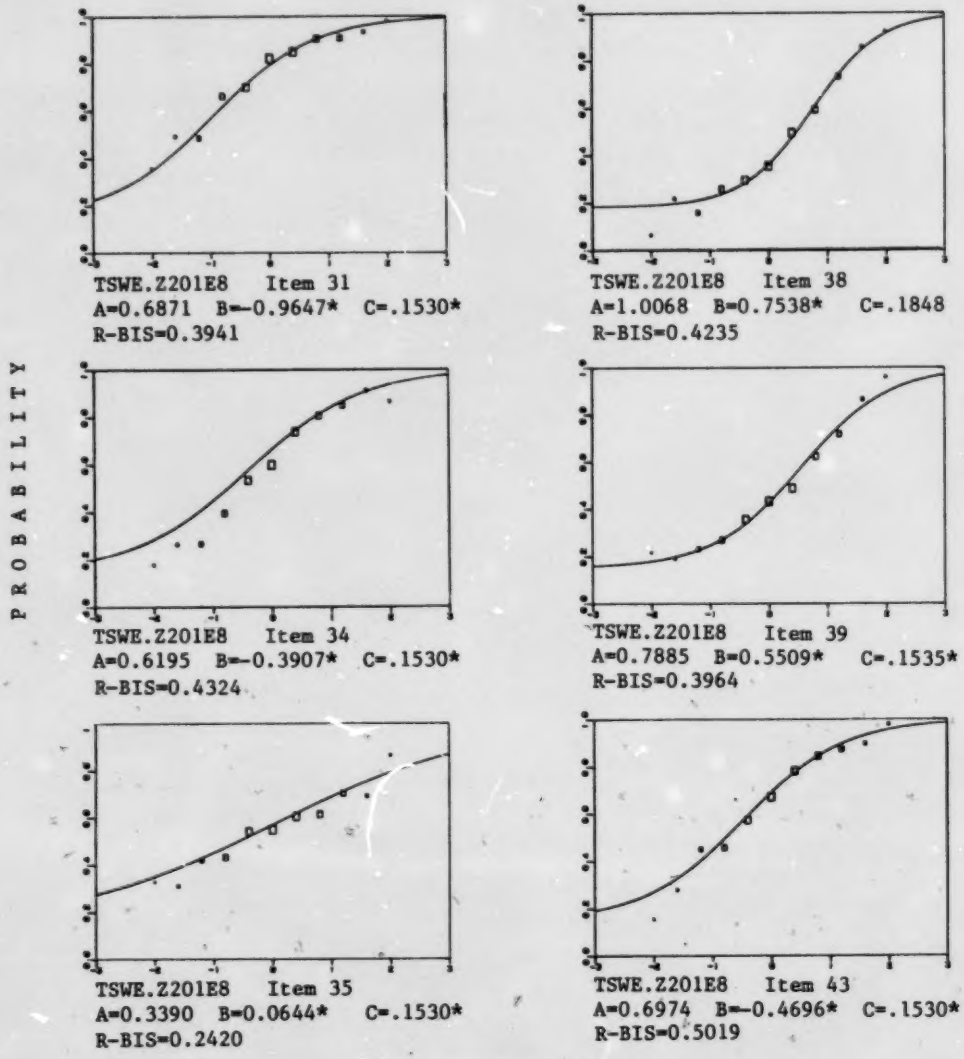


FIGURE 3 (Cont'd)

Evaluation of Fit at the Total Score Level

A logical extension of the previous procedure is to consider how well the model predicts the distribution of number-right scores in a given sample. Since the three-parameter model has no way of predicting omits, for a given individual the analysis is based on number-right score rather than formula scores.[1] The rationale of this procedure is to compare a prediction of the model—in this case the frequency distribution of number-right scores—against the empirical results—in this case the observed distribution of number-right scores. Although a number of indices could be used to quantify the discrepancies between the predicted and observed distribution, none was used since it would have required additional programming. Therefore, the assessment of fit will also be judgmental.

The observed frequency distribution of number-right scores was obtained by simply tabulating the number of testees at each number-right score level. The predicted frequency distribution was obtained by a complex algorithm; however, its conceptual equivalent is easily understood as follows:

For each testee determine n*, the number of items reached.

Compute the $P_i(\theta_a)$ and $Q_i(\theta_a)$, where $P_i(\theta_a)$ is the probability of answering the ith item correctly, as given by the three-parameter logistic model, for a given θ_a; $Q_i(\theta_a) = 1 - P_i(\theta_a)$.

Generate all possible $2n^*$ response vectors such that $u_i = 1$ indicates a correct response, and $u_i = 0$ indicates an incorrect response.

For each vector substitute $P_i(\theta_a)$ if $u_i = 1$ and $Q_i(\theta_a)$ if $u_i = 0$; multiply the probabilities to obtain the probability of the response vector. That is, compute:

$$\prod_{i=1}^{n^*} P_i(\hat{\theta}_a)^{u_i} \, Q_i(\hat{\theta}_a)^{(1 - u_i)}$$

Group response vectors with the same number of 1's. There are $n^* + 1$ such groups corresponding to number-right scores of 0, 1, 2, ... n^*.

Sum the probabilities of each response vector within a group. The sum of these probabilities is the expected frequency of this number-right score. When this is done for each group we have the expected distribution of number-right score, for one testee.

Repeat the above steps for each testee and sum the distribution over examinees for each number-right score.

Divide by N, the number of testees, which yields the expected distribution of number of right scores for the entire sample.

Notice that this procedure assumes local independence since we take the product of probabilities in the fourth step, which is the reason why the comparison against the observed distribution of number-right scores may be viewed as a test of fit.

---

1. The TSWE is scored operationally using scores corrected for guessing. We refer to such scores as formula scores.

TABLE 2. Mean and Standard Deviations for Observed and Expected Number-Right Distributions Including and Excluding Students with Omitted Items

Form Sample	Including Omits Mean	Including Omits S.D.	Excluding Omits Mean	Excluding Omits S.D.
X406E7				
Obs.	33.94	8.68	34.77	8.45
Exp.	34.03	8.95	34.72	8.82
N	2960	2960	2512	2512
Z101E7				
Obs.	31.26	8.69	32.17	9.49
Exp.	31.34	9.93	32.14	9.84
N	2973	2973	2461	2461
X506E8				
Obs.	32.86	8.43	33.63	8.19
Exp.	32.96	8.78	33.61	8.63
N	2980	2980	2514	2514
Z201E8				
Obs.	33.42	7.63	34.20	7.39
Exp.	33.58	7.94	34.18	7.81
N	2980	2980	2417	2417

The procedure was applied to the following data sets: X406E7, Z101E7, X506E8, and Z201E8, once excluding testees with omits and once including those students. The results are shown in Figure 4, but are best summarized by Table 2 which reports the mean and standard deviation of the observed and expected distribution. With omits included the expected mean and the expected standard deviation are somewhat larger than the corresponding observed values. With omits excluded the expected standard deviation is also larger but the expected mean now is slightly smaller than the observed value. Apart from these differences, however, the discrepancies between observed and expected means and standard deviations do not appear any larger for Z101E7 and Z201E8 where some of the b's had been fixed.

Factorial Structure of TSWE Data

The third method of assessing fit involves factor analysis. Attempts to examine fit through factor analysis (e.g., Indow and Samejima, 1966) have done on inter-item correlation matrices. By contrast, the present use of factor analysis involves correlation among subscores. Since the TSWE contains two item types, a reasonable hypothesis is that response to each item type requires somewhat different processes; that is, the two item types do not measure the same construct.

Method. Two formula scores were computed for each item type by totaling across odd and even items separately. To insure that the odd and the even scores were based on the same number of items, item 25 was excluded from the odd items for the usage items and item 40 was excluded from the even items for the sentence correction items.

Correlation and covariance matrices were computed based on the four scores for the following data sets: W506E4, X104E4, X106E5, X406E7, X506E8. The matrices appear in Appendix B. A two-factor model was fitted to each of these correlation matrices

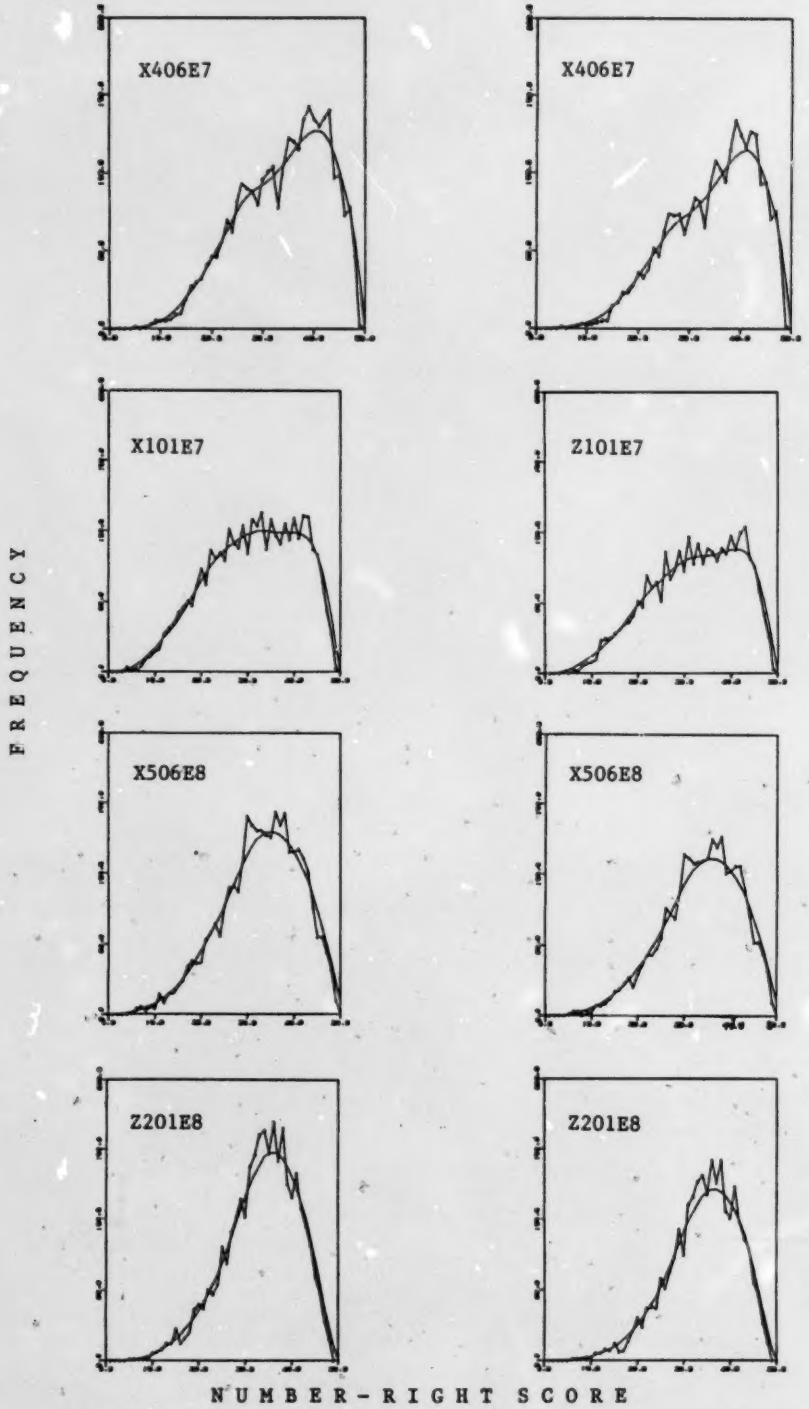


FIGURE 4. Observed and expected distribution of number-right score including (left) and excluding (right) students with omitted responses.

TABLE 3. Summary Results of Factor Analysis with Two Item Type Factors

Form	Sample	χ^2	df	p	Correlation Between True Scores*
W506	E3	.42	1	.52	.889 (.01)
X104	E4	2.77	1	.10	.884 (.01)
X106	E5	3.50	1	.06	.891 (.01)
X406	E7	.15	1	.70	.879 (.01)
X506	E8	18.04	1	.00	.915 (.01)

*The value in parenthesis is the asymptotic standard error of the correlation.

and the parameters estimated by the maximum likelihood procedure using the COFAMM Program (Sörbom and Jöreskog, 1976). The model tests the hypotheses that the correlation matrix, Σ, is described as follows:

$$\begin{bmatrix} x_1 & 0 \\ x_2 & 0 \\ 0 & x_3 \\ 0 & x_4 \end{bmatrix} \begin{bmatrix} 1 & x_5 \\ x_5 & 1 \end{bmatrix} \begin{bmatrix} x_1 & x_2 & 0 & 0 \\ 0 & 0 & x_3 & x_4 \end{bmatrix} + \begin{bmatrix} x_6 & 0 & 0 & 0 \\ 0 & x_7 & 0 & 0 \\ 0 & 0 & x_8 & 0 \\ 0 & 0 & 0 & x_9 \end{bmatrix}$$

The x's are parameters to be estimated. The 0's indicate the corresponding parameter is fixed to zero. x_1 through x_4 represent the factor loadings; x_5 is the correlation among the two factors each defined by an item type; and x_6 through x_9 represent the unique variance of each subscore. This model is discussed by Jöreskog (1978) who notes it is a restatement of an earlier model (1956) by Lord.

For our purposes the model can be used to test the presence of an item type effect by estimating the model with x_5 set to 1.0, that is, hypothesizing the correlation among true scores to be perfect. This was done for the five data sets mentioned earlier and, in every case, the hypothesis $x_5 = 1$ was rejected with $p < .0001$. The model was then estimated allowing x_5 to be estimated. The results are shown in Table 3. For all forms but E8 the model now fits ($p > .05$). Even when the p values are not accurate, since the data do not have a multivariate normal distribution as assumed by COFAMM, the magnitude of the chi-square statistic suggests that for E8 the two factor models did not account as well for all the correlations.

The estimated correlation between the true scores for the two item types is also shown in Table 3. The asymptotic standard error of the estimated correlation is shown in parenthesis. As can be seen the correlations are below .90, except for E8.

This analysis suggests that the structure of the TSWE can be understood by postulating two item-type factors. Although the correlation between the two scores is very high it does not approach 1.0 as would be expected if both scores measured a single

construct. More concretely removing the constraint that the two factors correlate 1.0 reduced the residuals to almost zero. To illustrate, the residual covariance matrices for W506E3 under the two models are shown below. The residuals under the hypothesis $x_5 = 1$ are shown above the diagonal. The corresponding residuals when x_5 is estimated are shown below the diagonal. (The first letter, E or O, stands for even and odd, respectively; US = usage, SC = sentence correction.)

	OUS	EUS	OSC	ESC
OUS	-	.012	-.016	-.014
EUS	.000	-	-.013	-.019
OSC	-.002	.002	-	.083
ESC	.002	-.002	.000	-

It can be seen all residuals are substantially reduced when x_5, the correlation among the two factors, is estimated rather than fixed to a value of 1.00. In short, the analysis presented here suggests that the internal structure of the TSWE can be understood better in terms of two item-type factors.

Summary

It was shown in this section that there appears to be some lack of fit of the three-parameter model to the TSWE data. This was obvious at the item level where some observed and expected items on ability regressions were not congruent at the subscore level where it appeared that two item type factors were necessary to fully account for the internal structure of the data, hence suggesting lack of undimensionality. Also, at the total score level, the model did not seem to reproduce the distribution of number-right scores completely accurately. The important question from a practical point of view is whether these deviations from the model have an impact on equating results. The next section presents the results relevant to that question.

ASSESSMENT OF PRE-EQUATING

The criterion for judging adequacy of pre-equating in the present study is by comparison to conventional equating. Implicit in this choice of criterion is the assumption that conventional equating provides a reasonable criterion. While this is not generally true, the conventional equating was done by spiraling the old and new forms in random samples of the new population. Furthermore, the test specifications are observed very strictly, and, as a result, there is minimal variation across forms of the test. All of this suggests conventional linear equating should work adequately with the TSWE. Nevertheless, three conventional equatings were used as criterion. One criterion, labeled C1, is the operational linear equating, that is, the procedure used in the reporting of scores. For E7 and E8 the operational equatings used SAT-V and SAT-M as anchor tests. The second criterion used, C2, was similar to C1 except only SAT-V was used as an anchor. Finally, the third criterion, C3, was equi-percentile equating using SAT-V as an anchor test. (It would have been desirable to use SAT-V and SAT-M as anchors for equi-percentile equating also, but the computer program did not allow it.) A description of linear and equi-percentile equating can be found in Petersen, Cook, and Stocking, 1981.

The comparison of pre-equating and operational equating results will be limited to two forms, E7 and E8, since only on these forms was it possible to simulate pre-equating conditions. For operational use, that is, to actually report converted scores to testees, E7 had been equated to E5 in the November 1975 administration; E8 had been also equated to E5 but in the December 1975 administration. In both cases, the "old" and "new" forms were spiraled, and scores on the SAT-V and SAT-M were used to adjust the TSWE scores before equating the TSWE means and standard deviations of the

two forms. The results of conventional linear equating are two parameters, usually referred to as A and B, which are used for converting formula scores to the TSWE metric as follows:

$$CS = A\,(FS) + B$$

where CS is the converted score and FS is the formula score. Since the TSWE has 50 items, FS ranges from -12 to 50. If CS is less than 20 it is set to 20. Also if CS becomes greater than 60 it is set to 60. The results of equal percentile equating is a table which converts scores in the old form to corresponding scores in the new form.

For methodological as well as practical reasons two levels of pre-equating were studied. The least demanding level consists of estimating IRT parameters for the new form, E7 or E8 in this case, when the items appear together as a form. Strictly speaking, this is not pre-equating but merely IRT-based true-score equating. We will refer to it as IRT-equating. (If the IRT parameters had been estimated on a different population it could be considered truly pre-equating). Nevertheless, precisely because it is a very undemanding form of pre-equating, the results from this comparison serve as a good benchmark to compare the results of pre-equating proper. The parameter sets for the new forms for this comparison were X406E7T and X506E8T for E7 and E8, respectively.

For the second level, pre-equating proper, the parameter sets E7P and E8P were used. For E7 and E8 the a, b and c's of 20 and 14 items, respectively, were taken from parameter sets X104P, X105P, and X106P. For the remaining items the a, b, and c's were taken from the parameter set Z100E7 for E7, and the parameter set Z201E8 for E8.

Within IRT-equating and pre-equating, three old forms were used, namely E3, E4, and E5, to put the new form E7 or E8 on the TSWE scale.

Equating Based on IRT

The procedure used to transform formula scores on the new form to scaled scores can be described in general as follows: For a given true score, on the new form find the corresponding θ. Next, find the true score on the "old" form associated with this θ. Finally, apply existing conversion parameters to put the equated true scores on the TSWE scale. Since the TSWE is scored using formula scores, the actual procedure is based on true formula scores. A step-by-step description follows.

For each integer formula score on the new form FS_{new}, greater than

$$\ell = \sum_{i=1}^{50} c_{i,new}$$

and less than or equal to 50 (where $c_{i,new}$ is the c estimated parameter for the ith item) compute the associated true score number right scale as follows

$$NR_{new} = .80\ FS_{new} + 50/5$$

This is based on the fact that if an examinee attempts all items in the test, number-right and formula score are linearly related with slope m/(m-1), where m is the number of alternatives, and constant n/(m-1), where n is the number of items in the test. A similar relationship holds for true formula scores and true scores as shown by Lord (1980, Chapter 15).

The next step is to find the θ associated with a given NR_{new}. This is done by solving for θ in the equation:

$$NR_{new} = \sum_{i=1}^{50} P_i^*(\theta)$$

where the $P_i^*(\theta)$ is computed using the $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$, for the new form.

Having found the needed θ compute the corresponding true score in the old form as follows:

$$NR_{old} = \sum_{i=1}^{50} P_i(\theta)$$

where now $P_i(\theta)$ is computed using the $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$ from the old form.

The true formula score corresponding to this true score is

$$FS_{old} = NR_{old}/.80 - 50/4$$

Finally, FS_{old} is converted by means of existing parameters A and B as follows:

$$CS = A(FS_{old}) + B$$

If FS_{old} is less than ℓ a somewhat different procedure is used.

The procedure is described by Appendix C of Chapter 13 of Lord (1980). This procedure was applied with E3, E4, and E5 as old forms (using parameter sets W506E3, X104,5E4T, and X106E5T, respectively) and new forms E7 and E8 (using parameter sets X406E7T and X506E8T for IRT-equating, and parameter sets E7P and E8P for pre-equating).

Results

The results of the three criterion equatings were all very close. This can be seen in Tables 4 and 5 which report the criterion equatings corresponding to every five formula raw score points. (Full point-by-point conversions are found in Appendix C.) The largest discrepancy among criterion is 1 point for both E7 and E8. The discrepancy of IRT-equating and pre-equating with the criterion equatings, however, is larger.

The magnitude of the discrepancies can be appreciated by examining the mean and standard deviation corresponding to the criterion equatings, IRT-equating and pre-equating. The mean and standard deviations are based on the frequency distribution observed on the first national administration of E7 and E8. (These frequencies can be found in Appendix C.) In particular, four trends are more or less obvious. First, operational and IRT-based equatings are much more discrepant for E8 than for E7. Secondly, for the IRT-based conversions the mean is higher and the standard deviation smaller compared to the criterion equatings. Thirdly, the choice of an old form seems to affect the discrepancy of IRT-based and operational equating. More concretely, using E3 as an old form for either pre-equating or IRT-equating yields the most discrepant results. Finally, comparing the results for pre-equating and IRT-equating, pre-equating for E7 actually yields less discrepant results than IRT-equating but the opposite is true for E8.

A more detailed analysis of the results can be obtained from an index suggested by Marco, Petersen, and Stewart (1979), namely, the weighted mean squared error,

$$\Sigma_j f_j d_j^2/N = \Sigma f_j (d_j - \bar{d})^2/N + \bar{d}^2 \quad \text{or}$$

(total error) = (variance of differences) + (squared bias)

TABLE 4. Summary Conversion Table Comparing Conventional Equating, IRT-equating, and Pre-equating for E7

Raw Score	Criterion C1	Criterion C2	Criterion C3	IRT-equating Old Form E3	IRT-equating Old Form E4	IRT-equating Old Form E5	Pre-equating Old Form E3	Pre-equating Old Form E4	Pre-equating Old Form E5
50	60	60	60	60.	60.	60.	60.	60.	60.
45	58	58	59	58.	58.	59.	57.	57.	58.
40	53	53	53	54.	54.	54.	53.	52.	52.
35	48	48	48	49.	49.	49.	48.	48.	48.
30	43	43	43	45.	44.	44.	44.	44.	43.
25	38	38	38	40.	39.	39.	40.	39.	39.
20	33	33	33	35.	34.	34.	35.	34.	34.
15	28	28	28	30.	28.	28.	30.	29.	29.
10	22	22	22	24.	23.	23.	25.	24.	24.
5	20	20	20	20.	20.	20.	20.	20.	20.
0	20	20	20	20.	20.	20.	20.	20.	20.
-5	20	20	20	20.	20.	20.	20.	20.	20.
-10	20	20	20	20.	20.	20.	20.	20.	20.
Mean	43.84	43.83	43.69	45.25	44.55	44.78	44.66	44.06	44.07
S.D.	9.70	9.73	9.80	9.20	9.64	9.81	8.62	8.93	9.05

C1 is based on linear observed score equating using SAT-V and SAT-M as anchors; C2 only uses SAT-V as an anchor; C3 is based on equi-percentile equating. The means and standard deviations are based on the formula score frequency distribution for the first national administration of E7.

TABLE 5. Summary Conversion Table Comparing Conventional Equating, IRT-equating, and Pre-equating for E8

Raw Score	Criterion C1	Criterion C2	Criterion C3	IRT-equating Old Form E3	IRT-equating Old Form E4	IRT-equating Old Form E5	Pre-equating Old Form E3	Pre-equating Old Form E4	Pre-equating Old Form E5
50	60	60	60	60.	60.	60.	60.	60.	60.
45	59	59	60	58.	59.	60.	58.	58.	59.
40	53	53	54	53.	53.	53.	54.	53.	54.
35	47	48	48	49.	48.	48.	49.	49.	48.
30	42	42	42	43.	43.	42.	44.	43.	43.
25	36	36	35	38.	37.	37.	39.	38.	38.
20	30	30	29	33.	31.	32.	34	33	33.
15	24	24	24	27.	26.	26.	29.	27.	28.
10	20	20	20	22.	21.	21.	23.	22.	22.
5	20	20	20	20.	20.	20.	20.	20.	20.
0	20	20	20	20.	20.	20.	20.	20.	20.
-5	20	20	20	20.	20.	20.	20.	20.	20.
-10	20	20	20	20.	20.	20.	20.	20.	20.
Mean	42.10	42.14	42.04	43.59	42.84	42.94	44.30	43.60	43.59
S.D.	9.96	9.98	10.21	8.77	9.26	9.31	8.34	8.78	8.77

See footnote to Table 4. The means and standard deviations are based on the formula score frequency distribution for the first national administration of E8.

where $d_j = (t'_j - t_j)$, t'_j is the criterion score (which in this case corresponds to the operational score) for raw score x_j; t_j is the IRT-based converted score corresponding to the same raw score x_j; $\bar{d} = f_j\ d_j/N\ f_j$ is the frequency of x_j and $N = \Sigma f_j$.

Tables 6 and 7 show the computed indices for E7 and E8, respectively. The f's used correspond to the frequency of x in the first national administration and can be found in Appendix C.

An examination of the discrepancy indices largely corroborates the results noted earlier. Within a given equating procedure there is variation due to the choice of old form. For IRT-equating E3 yields the most discrepant results in terms of the weighted mean squared differences, E4 the least discrepant results and E5 is in between. This is true for both E7 and E8. For pre-equating E3 also yields the most discrepant results but E5 yield the least discrepant results with E4 in between. Again, this is true for both E7 and E8.

For E7, C1, the linear equating using SAT-V and SAT-M as anchors, yields the least discrepant results for both IRT-equating and pre-equating followed by C2, linear equating using only SAT-V as anchor, and C3, equal percentile equating. For E8, however, using linear equating with only SAT-V as anchor yields the least discrepant results followed by C1 and C3 in that order. That is, using equi-percentile equating as a criterion yielded the most discrepant results for both E7 and E8 and IRT-equating and pre-equating.

Comparing IRT-equating and pre-equating it can be seen from Table 6 that for E7 pre-equating is actually closer to the criterion equatings but that the composition of the mean squared error is different. For IRT-equating the squared bias is the larger component whereas for pre-equating the variance of the differences is actually the larger component. For E8, however, the square bias is the larger component for both IRT-equating and pre-equating.

SUMMARY AND CONCLUSIONS

This investigation was concerned with how well IRT equating and pre-equating could reproduce the conversion line for two TSWE forms which had been previously equated by conventional observed score equating methods. The approach was to determine first how well TSWE data fitted the three-parameter logistic model and then to compare IRT equating and pre-equating against three criterion equatings so that discrepancies in equating could be traced to more fundamental questions of fit.

The various procedures for investigating fit suggested several violations of the assumptions of the model. At the item level some of the estimated item-on-ability regressions did not fit the data as well when the b parameter had been fixed to its estimated value based on a pretest administration. This is important in pre-equating since presumably in practical application parameter estimates would be obtained from pretests. However, the fact that the problem was observed on just a few items suggest that the problem may not be too serious.

At the subscore level it was shown that two factors, corresponding to the two TSWE item types, were required to account for the internal structure of the TSWE, thus suggesting a violation of the undimensionality assumption. This is to be expected, and perhaps, so long as the nature of multidimensionality is constant across sample-form combinations, no great harm would occur. It so happened, however, that for form E8 the two-factor model that fitted the other forms did not fit as well. Furthermore, the equating parameters derived under conventional procedures are very

TABLE 6. Weighted Mean Square Difference for Form E7, Using E3, E4, and E5 as the "Old" Form and Three Different Criteria

		IRT-equating Old form			Pre-equating Old Form		
		E3	E4	E5	E3	E4	E5
Mean squared difference criterion	C1	2.53	.71	.94	2.04	.79	.70
	C2	2.57	.73	.96	2.11	.84	.74
	C3	3.39	1.33	1.43	2.77	1.19	.88
Variance of difference criterion	C1	.53	.20	.89	1.37	.74	.64
	C2	.53	.20	.92	1.42	.79	.68
	C3	.95	.58	1.19	1.83	1.06	.73
Squared bias criterion	C1	2.00	.51	.05	.67	.05	.06
	C2	2.04	.53	.04	.69	.05	.06
	C3	2.44	.74	.24	.94	.13	.15

C1 is based on linear observed score equating using SAT-V and SAT-M as anchors; C2 only uses SAT-V as an anchor; C3 is based on equi-percentile equating. The weighting function is the formula score frequency distribution for the first national administration of E7.

TABLE 7. Weighted Mean Squared Difference for Form E8, Using E3, E4, and E5 as the "Old" Form and Three Different Criteria

		IRT-equating Old form			Pre-equating Old form		
		E3	E4	E5	E3	E4	E5
Mean squared difference criterion	C1	3.99	1.22	1.34	7.67	4.00	3.81
	C2	3.84	1.17	1.29	7.53	3.85	3.76
	C3	4.89	1.70	1.89	8.81	4.89	4.68
Variance of difference criterion	C1	1.75	.66	.63	2.83	1.73	1.58
	C2	1.75	.69	.66	2.90	1.73	1.67
	C3	2.49	1.07	1.09	3.74	2.48	2.30
Squared bias criterion	C1	2.24	.55	.71	4.84	2.27	2.23
	C2	2.10	.48	.63	4.64	2.12	2.09
	C3	2.39	.63	.80	5.07	2.42	2.38

See footnote to Table 6. The weighting function is the formula score frequency distribution of the first national administration of E8.

different for E8 compared to the other forms. This suggests that the internal structure of E8 was somewhat different.

Finally, at the total score level, a slight bias was observed in the prediction of the mean and standard deviation of number right scores. The direction of the bias, however, depended on whether students with omitted responses were excluded or included in the data. When students with omitted responses were included in the data the mean and standard deviations of the expected distribution were slightly larger than the corresponding observed values. When students with omitted items were excluded the mean of the expected distribution was slightly smaller than the observed mean but the standard deviation was still larger than the observed value.

The bias in the mean when students with omits are included appears to be partly due to the fact that the omitted responses are not counted at all in the observed data whereas the calculation of the expected distribution assumes all items are answered. Thus if students were instructed to respond to the omitted items their number-right score would increase. In fact, when students with omits are excluded the bias changes direction, although it is of a very small magnitude, on the order of .02 or .03. There is no obvious explanation for this "net" bias in the mean or the bias in the standard deviation. It is not clear, for example, whether the bias is due to the apparent multidimensionality of the data or an inherent bias in the LOGIST procedure. Clearly, further research is needed.

Departures from the models are to be expected with actual data. The important question, and the focus of this study, is whether such departures seriously affect equating. To answer this question IRT equating and pre-equating were done for TSWE forms E7 and E8. The results for E8 were disappointing in that large discrepancies were found between the operational conversion line and the IRT-based equatings. However, since the two-factor model that fitted all forms did not fit E8, this appears to be the result of the aberrant internal structure of E8 rather than a failure of IRT equating. In other words it is not clear that E8 is properly equated even by conventional methods, and, hence, for E8 the converted scores may not be a good criterion. Therefore, it is wiser to formulate conclusions based on the results for E7 only.

The equating results for E7 were much more favorable, but some consistent discrepancies were observed including an overestimation of the mean as well as an underestimation of the standard deviation of the distribution of converted scores. The overestimation of the converted score mean would seem to be consistent with the fact that the mean number-right score is also overestimated when students with omits are included in the data.

The underestimation of the standard deviation of converted scores, however, is inconsistent with the earlier finding that the standard deviation of number-right scores is actually overestimated by the IRT model. This suggests that something in the transformation of formula scores to scaled scores is responsible for the misrepresentation of the standard deviation. At least two limitations of the transformation procedure are obvious. One is the assumption of a linear relationship between formula scores and number-right scores, which is false if students with omitted responses are included. A second limitation of the procedure lies in the fact that to put the new form on scale it is necessary to apply A and B conversion parameters based on observed score equating to true formula scores. Unfortunately, it is not obvious what alternative procedure could be devised to put on scale new forms so long as formula scoring was in use. This suggests that IRT-equating and pre-equating would be more accurate if number-right scoring was used rather than formula scoring. Number-right scoring is no panacea, however, so long as tests are speeded (and they always will be for at least some students under the usual administration procedures). The problem is that under number-right scoring it is advantageous to the student to attempt an item even if he or she has to guess. If they actually do so, they will not be responding as a function of their ability and will thus create a violation of the model (Lord, 1980).

As for pre-equating, based on E7, there is reason to be optimistic since the mean squared differences were not consistently higher for pre-equating across old forms or criteria. However, the criteria for evaluating both IRT-equating and pre-equating are not defensible on other than practical grounds. Thus, unless the comparability of IRT and pre-equating were to change when evaluated against a more adequate criteria, we can reasonably expect that as better procedures for linking parameters are developed (see e.g., Petersen et al., 1981) pre-equating will prove to be a feasible operational procedure.

REFERENCES

Angoff, W.H., "Scales, Norms, and Equivalent Scores," in Educational Measurement (2nd ed.), R.L. Thorndike, editor. Washington, D. C.: American Council on Education, 1971.

Birnbaum, A., "Test Scores, Sufficient Statistics, and the Information Structure of Tests," in Statistical Theories of Mental Test Scores, F.M. Lord and M.R. Novick, editors. Reading, Mass.: Addison-Wesley, 1968.

Breland, H.M., A Study of College English Placement and the Test of Standard Written English. Project Report 77-1. Princeton, N.J.: Educational Testing Service, 1976.

Indow, T., and F. Samejima, On the Results Obtained by the Absolute Scaling Model and the Lord Model of the Field of Intelligence. Yokohama: Psychological Laboratory, Hiyoshi Campus, Keio University, 1966.

Jöreskog, K.G., "Structural Analysis of Covariance and Correlation Matrices," Psychometrika, Vol. 43, 1978, pp. 443-477.

Lord, F.M., Applications of Item Response Theory to Practical Testing Problems. In press, 1980.

Marco, G.L., N.S., Petersen, and E.E., Stewart, A Test of Adequacy of Curvilinear Score Equating Models. Paper presented at the 1979 Computerized Adaptive Testing Conference, Minneapolis, June 1979.

Petersen, N.S., L.L. Cook, and M.L. Stocking, IRT Versus Conventional Equating Methods: A Comparative Study of Scale Stability. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, 1981.

Sörbom, D., and K.G. Jöreskog, COFAMM: Confirmatory Factor Analysis with Model Modification, a FORTRAN IV Program. National Educational Resources, 1976.

Wood, R.L., M.S. Wingersky, and F.M. Lord, LOGIST—A Computer Program for Estimating Examinee Ability and Item Characteristic Curve Parameters. Research memorandum 76-6. Princeton, N.J.: Educational Testing Service, 1976.

Yen, W.M., "The Extent, Causes, and Importance of Context Effects on Item Parameters for Two Latent Trait Models," Journal of Educational Measurement, Vol. 17, 1980, pp. 297-311.

APPENDIX A: Transformation of the b 's to Put the LOGIST Output from a New Administration on the Same Scale as the Output from an Old Administration

Robust estimates of scale and location are used to determine the slope and intercept of the line relating the b's estimated from two samples of examinees.

The estimate of scale for each of the forms is a biweight estimate (see Mosteller and Tukey, 1977). The formulas are

$$\tilde{b} = \text{median of b's},$$

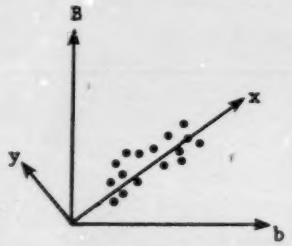
$$u_i = \frac{b_i - \tilde{b}}{9(MAD_b)}$$

where MAD_b = Median absolute deviation = median $|b_i - \tilde{b}|$,

$$s_b^2 = \frac{\sum_i' (b_i - \tilde{b})^2 (1 - u_i^2)^4}{[\sum_i' (1 - u_i^2)(1 - 5u_i^2)][\sum_i' (1 - u_i^2)(1 - 5u_i^2) - 1]}$$

where $\sum_i'$ indicates summation for $u_i^2 \leq 1$.

Let B_i be the b's on the new sample and b_i be the b's on the old sample.



The slope of the line relating the b's is taken to be

$$m = s_B / s_b .$$

Define an xy coordinate system by

$$x = (b + mB),$$

$$y = (-mb + B).$$

Get a robust estimate of location separately for x and for y using the formulas on page 205 of Mosteller and Tukey. Let y* = median of the y's,

$$w_i = \begin{cases} (1 - (\frac{y_i - y^*}{c(MAD_y)})^2)^2 & \text{when } (\frac{y_i - y^*}{c(MAD_y)})^2 < 1 \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

where MAD_y = Median $|y_i - y^*|$, c = 6. Compute a new y* from the formula

$$y^* = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \qquad (2)$$

Iterate through equations 1 and 2 until the change between two estimates of y* is less than .0001. Repeat the process for the x's.

Transform these biweight estimates of location in the xy coordinates back to the bB coordinates and require that the line with slope m pass through this point

$$B^* = (mx^* + y^*) / (m^2 + 1) ,$$

$$b^* = (x^* - my^*) / (m^2 + 1) .$$

The equation for the line that puts the old parameters on the new parameter scale is

$$b^T = mb + B^* - mb^* \qquad (3)$$

and

$$a^T = (1/m)a. \qquad (4)$$

The equation for the line that transforms the new parameters to the old parameter scale is

$$B^T = \frac{B}{m} - \left[\frac{B^* - mb^*}{m} \right] \qquad (5)$$

and

$$A^T = mA \qquad (6)$$

where a and A are the discrimination estimates based on the old and new samples, respectively.

To put the parameters of a new form onto the same scale as an old form, the new form and the old form must be administered to random samples in a new administration. This was done by spiraling. For the new administration we reestimate the parameters for the old form and estimate the parameters for the new form or random samples of equal size. The new form is put onto the scale of the old form in the new administration by setting the means and standard deviations of the abilities equal. This is done in LOGIST by standardizing the abilities to a mean of 0 and a standard deviation of 1 for both forms. Then the transformation that puts the old form, new sample, equations 5 and 6, onto scale is applied to the parameters for the new form to put those parameters onto scale.

Reference

Mosteller, F., and J.W. Tukey, Data Analysis and Regression. Reading: Mass.: Addison-Wesley, 1977.

APPENDIX B: Variance-Covariance, Correlation Matrices, and Mean Vectors for Five TSWE Forms

Form		EUS	OUS	ESC	OSC	Means
W506E3	EUS	17.3214	12.4191	4.8800	4.1147	9.6094
	OUS	.7535	15.6848	4.6327	3.8515	10.4343
	ESC	.5842	.5828	4.0280	1.9298	4.3198
	OSC	.5556	.5466	.5404	3.1659	4.4243
X104E4	EUS	15.9720	12.5510	4.5698	4.5351	9.3356
	OUS	.7488	17.5905	4.9187	4.7088	9.1504
	ESC	.5665	.5810	4.0743	2.2352	4.2981
	OSC	.5639	.5579	.5503	4.0494	3.3263
Y106E5	EUS	16.6073	11.5705	4.2797	4.7738	9.0585
	OUS	.7210	15.5071	4.2672	4.5542	9.5068
	ESC	.5218	.5384	4.0514	2.1700	3.5152
	OSC	.5873	.5798	.5405	3.9760	4.4588
X406E7	EUS	16.1853	12.5826	4.4694	3.8127	9.6603
	OUS	.7574	17.0511	4.8645	4.1165	11.0141
	ESC	.5999	.6361	3.4295	1.9011	4.7617
	OSC	.5236	.5507	.5671	3.2765	4.8550
X506E8	EUS	15.0591	9.8126	4.0468	4.3847	9.4296
	OUS	.6888	13.4784	3.7155	3.6014	10.3031
	ESC	.5678	.5511	3.3727	1.8783	4.6323
	OSC	.5682	.4933	.5143	3.9549	4.0061

Notation: EUS, even usage subscore; OUS, odd usage subscore; ESC, even sentence correction subscore; and OSC, odd sentence correction subscore. The correlation matrix appears below the diagonal.

APPENDIX C

TABLE C1. Point-by-Point Conversion Tables for Form E7

Raw Score	Criterion C1	Criterion C2	Criterion C3	IRT-equating Old Form E3	IRT-equating Old Form E4	IRT-equating Old Form E5	Pre-equating Old Form E3	Pre-equating Old Form E4	Pre-equating Old Form E5	Frequency
50-47	60	60	60	60	60	60	59	59	60	35
46	59	59	60	59	59	60	58	58	59	935
45	58	58	59	58	58	59	57	57	58	1091
44	57	57	57	57	57	58	56	56	57	1215
43	56	56	56	56	56	57	55	55	55	1286
42	55	55	55	55	55	56	54	54	54	127
41	54	54	54	55	54	55	53	53	53	1274
40	53	53	53	54	54	54	53	52	52	1362
39	52	52	51	53	53	53	52	51	51	1378
38	51	51	50	52	52	52	51	51	50	1402
37	50	50	50	51	51	51	50	50	50	224
36	49	49	49	50	50	50	49	49	49	1336
35	48	48	48	49	49	49	48	48	48	1390
34	47	47	46	49	48	48	48	47	47	1374
33	46	46	45	48	47	47	47	46	46	1369
32	45	45	45	47	46	46	46	45	45	327
31	44	44	44	46	45	45	45	44	44	1254
30	43	43	43	45	44	44	44	44	43	1274
29	42	42	42	44	43	43	43	43	43	1252
28	41	41	41	43	42	42	43	42	42	1209
27	40	40	40	42	41	41	42	41	41	370
26	39	39	39	41	40	40	41	40	40	1098
25	38	38	38	40	39	39	40	39	39	1081
24	37	37	37	39	38	38	39	38	38	1045
23	36	36	36	38	37	37	38	37	37	967
22	35	35	35	37	36	36	37	36	36	352
21	34	34	34	36	35	35	36	35	35	845
20	33	33	33	35	34	34	35	34	34	804
19	32	32	32	34	33	33	34	33	33	750
18	31	31	31	33	31	32	33	32	32	679
17	30	30	30	32	30	31	32	31	31	276
16	29	29	29	31	29	30	31	30	30	573
15	28	28	28	30	28	28	30	29	29	516
14	27	27	26	28	27	27	29	28	28	463
13	26	26	25	27	26	26	28	27	27	397
12	25	24	24	26	25	25	27	26	26	166
11	24	23	23	25	24	24	26	25	25	300
10	22	22	22	24	23	23	25	24	24	258
9	21	21	21	23	22	22	24	23	23	228
-12-8	20	20	20	22	21	21	23	22	22	179

Note: The frequencies shown are based on the first national administration of E7 (November 1975) except they have been divided by 10. The criterion equatings are based on observed score equating methodology as described in the text. C1 is based on linear observed score equating using SAT-V and SAT-M as anchors; C2 only uses SAT-V as anchor; C3 is based on equal percentile equating.

TABLE C2. Point-by-Point Conversion Table for Form E8

Raw Score	Criterion C1	Criterion C2	Criterion C3	Criterion C4	IRT-equating Old Form E3	IRT-equating Old Form E4	IRT-equating Old Form E5	Pre-equating Old Form E3	Pre-equating Old Form E4	Pre-equating Old Form E5	Frequency
46-50	60	60	60	60	59	60	60	59	59	60	251
45	59	59	60	59	58	59	60	58	58	59	360
44	58	58	58	58	57	57	58	57	57	58	463
43	57	57	[illegible]	57	56	56	57	56	56	57	537
42	56	56	[illegible]	55	55	55	56	55	55	56	59
41	55	55	55	54	54	54	54	54	54	55	604
40	53	53	54	53	53	53	53	54	53	54	679
39	52	52	52	52	52	52	52	53	52	52	749
38	51	51	51	51	51	51	51	52	51	51	819
37	50	50	50	49	50	50	50	51	51	50	127
36	49	49	49	48	50	49	49	50	50	49	821
35	47	48	48	47	49	48	48	49	49	48	854
34	46	46	46	46	48	47	47	48	48	47	885
33	45	45	45	45	47	46	46	47	47	46	880
32	44	44	44	43	46	45	45	46	46	45	210
31	43	43	43	42	45	44	44	45	45	44	851
30	42	42	42	41	43	43	42	44	43	43	844
29	40	40	40	40	42	41	41	43	42	42	832
28	39	39	39	39	41	40	40	42	41	41	763
27	38	38	38	38	40	39	39	41	40	40	243
26	37	37	37	36	39	38	38	40	39	39	728
25	36	36	35	35	38	37	37	39	38	38	706
24	35	35	34	34	37	36	36	38	37	37	652
23	33	33	33	33	36	35	35	37	36	36	616
22	32	32	32	32	35	33	34	36	35	35	225
21	31	31	30	30	34	32	33	35	34	34	519
20	30	30	29	29	33	31	32	34	33	33	491
19	29	29	28	28	32	30	30	33	32	32	444
18	27	27	27	27	30	29	29	32	30	31	401
17	26	26	26	26	29	28	28	31	29	30	173
16	25	25	25	24	28	27	27	30	28	29	321
15	24	24	24	23	27	26	26	29	27	28	287
14	23	23	23	22	26	25	25	28	26	27	274
13	22	22	22	21	25	24	24	26	25	25	223
-12-12	20	20	20	20	24	23	23	25	24	24	103

Note: The frequencies shown are based on the first national administration of E8 (December 1975) except they have been divided by 10. The criterion equatings are based on observed score equating methodology as described in the text. C1 is based on linear observed score equating using SAT-V and SAT-M as anchors; C2 only uses SAT-V as anchor; C3 is based on equal percentile equating. C4 was derived by the same procedure used for C1 but using a different "old" sample.